IMPROVED RELATIVE DISCRIMINATIVE CRITERION USING RARE AND INFORMATIVE TERMS AND RINGED SEAL SEARCH-SUPPORT VECTOR MACHINE TECHNIQUES FOR TEXT CLASSIFICATION

WAREESA SHARIF

A thesis submitted in

fulfilment of the requirement for the award of the

Degree of Doctor of Philosophy

Faculty of Computer Science and Information Technology

Universiti Tun Hussein Onn Malaysia

DECEMBER, 2019

DEDICATION

I dedicate this study to my family

Dr Muhammad Ashraf,

Muhammad Mahad Ashraf,

Muhmmad Shawal Ashraf.

ACKNOWLEDGEMENT

Surely All praise is for Allah Almighty Who is the creator of this universe and Darood and Salam upon the Holy prophet Hazrat Muhammad (PBUH). I want to thank ALLAH Almighty for enabling me the strength and courage to complete Doctoral research. He has been the biggest source of strength for me. I would like to express my deepest gratitude to my Supervisors Dr Noor Azah Samsudin and Prof. Dr Mustafa Mat Deris. It's honor to me to be their PhD student. I am grateful for all their contributions of time, ideas, and knowledge to make my research experience productive and exciting.

I am thankful to my beloved husband Dr Muhammad Ashraf who encouraged me to take admission in PHD. His guidance and instructions were particularly valuable during preparing research papers and the thesis. I am also special thankful to my dearest sons Muhammad Mahad Ashraf and Muhammad Shawal Ashraf for their patience. I would like to thank University Tun Hussein Onn Malaysia (UTHM) for financial support under Graduate Researcher Incentive Grant (GIPS) Vote No. U497.

I am grateful to Amir, Abdullah, Iwan Tri Riyadi and Rashid who shared my interests and helped me to clarify my views through conversations and implementation. I would like to show my appreciation and pleasure to my family members, Rana Muhammad Sharif, Raisa, Anwaar, Ikram, Sajjad, Ali, Abdurehman, Samina, Aneesa, Amna and Dr Maria. I am also thankful to my loving friends Ashikin, Nor Hanifa from Malaysia. I would like to thank to family for pray, love and encouragement. The moral support of all my family members was the main stream to complete my PhD.

Thanks Malaysia.

Wareesa Sharif

ABSTRACT

Classification has become an important task for automatically classifying the documents to their respective categories. For text classification, feature selection techniques are normally used to identify important features and to remove irrelevant, and noisy features for minimizing the dimensionality of feature space. These techniques are expected particularly to improve efficiency, accuracy, and comprehensibility of the classification models in text labeling problems. Most of the feature selection techniques utilize document and term frequencies to rank a term. Existing feature selection techniques (e.g. RDC, NRDC) consider frequently occurring terms and ignore rarely occurring terms count in a class. However, this study proposes the Improved Relative Discriminative Criterion (IRDC) technique which considers rarely occurring terms count. It is argued that rarely occurring terms count are also meaningful and important as frequently occurring terms in a class. The proposed IRDC is compared to the most recent feature selection techniques RDC and NRDC. The results reveal significant improvement by the proposed IRDC technique for feature selection in terms of precision 27%, recall 30%, macro-average 35% and micro- average 30%. Additionally, this study also proposes a hybrid algorithm named: Ringed Seal Search-Support Vector Machine (RSS-SVM) to improve the generalization and learning capability of the SVM. The proposed RSS-SVM optimizes kernel and penalty parameter with the help of RSS algorithm. The proposed RSS-SVM is compared to the most recent techniques GA-SVM and CS-SVM. The results show significant improvement by the proposed RSS-SVM for classification in terms of accuracy 18.8%, recall 15.68%, precision 15.62% and specificity 13.69%. In conclusion, the proposed IRDC has shown better performance as compare to existing techniques because its capability in considering rare and informative terms. Additionally, the proposed RSS- SVM has shown better performance as compare to existing techniques because it has capability to improve balance between exploration and exploitation.

ABSTRAK

Pengkelasan merupakan suatu tugas penting untuk mengkelaskan dokumen kepada kategori tertentu. Untuk pengkelasan teks yang tepat, teknik pemilihan ciri biasanya diguna untuk mengenalpasti ciri penting dan menyingkirkan ciri yang tidak dikehendaki dan hingar. Teknik tersebut bertujuan meningkatkan kecekapan, ketepatan dan kebolehfahaman model pengkelasan dalam pelabelan teks. Kebanyakan teknik pemilihan ciri menggunakan kekerapan dokumen dan kekerapan istilah untuk menentukan kedudukan suatu istilah. Berbeza daripada kekerapan dokumen, kekerapan istilah menyokong nilai sebenar suatu istilah. Teknik pemilihan ciri sedia ada seperti *Relative Discriminant Criterion (RDC*) mengambil kira istilah kerap berlaku dan tidak mengendahkan istilah jarang berlaku dalam suatu kelas. Oleh yang demikian, berlaku pnurunan nilai-F apabila bilangan ciri bertambah. Penyelidikan ini mencadangkan teknik *Improved Relative Discriminative Criterion (IRDC)* yang mengambil kira istilah jarang berlaku kerana istilah jarang berlaku juga tidak kurang pentingnya berbanding kekerapan istilah berlaku dalam suatu kelas. *IRDC* dibanding kan dengan teknik *RDC* dan *Normalized RDC (NRDC).* Eksperimen dengan data Reuter21578, 20newsgroup dan TDT2, menunjukkan *IRDC* mencapai keputusan 27% ketepatan, 30% pemulihan, 35% purata-makro dan 30% purata-mikro. Tambahan pula, teknik *Ringed Seal Search-Support Vector Machine (RSS-SVM)* yang dicadang turut meningkatkan keupayaan membuat generalisasi dan pembelajaran dengan mengoptimumkan parameter kernel dan penalti. *RSS-SVM* dibandingkan dengan *Genetic Algorihm-Support Vector Machine (GA-SVM)* dan *Cuckoo Search (CS-SVM).* Eksperimen dengan set data Reuter21578, 20newsgroup dan TDT2, menunjukkan *RSS-SVM* mampu mencapai ketepatan 18.8%, pemulihan 15.68%, ketepatan 15.62% dan kekhususan 13.69%. Kesimpulannya, teknik *IRDC* yang dicadng kan telah menunjukkan prestasi lebih baik berbanding *RDC* dan *NRDC* dengan mengambil kira istilah yang jarang berlaku dan informatif. Manakala, *RSS-*

SVM menunjukkkan prestasi lebih baik berbanding GA-SVM dan CS-SVM apabila mampu menyeimbangkan antara penerokaan dan eksploitasi.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF SYMBOLS AND ABBREVIATIONS

TPR_{tc}	–	True Positive Rate (IRDC)
FPR_{tc}	–	False Positive Rate (IRDC)
tpr_{tc}	–	true positive rate (RDC)
fpr_{tc}	–	false positive rate (RDC)
ϵ	–	Small Number
ξ; ϕ $(\cdot)$	–	Error Term
C	–	Penalty Parameter
γ	–	Kernel parameter
ω	–	External noise
TP	–	*True Positive*
FP	–	*False Positive*
IRDC Criterion	–	Improved Relative Discriminative
RDC	–	Relative Discriminative Criterion
NRDC Criterion	–	Normalized Relative Discriminative
TC	–	Text Classification
UB	–	Upper bound
LB	–	Lower bound
K	–	Standard deviation
RSS	–	Ringed Seal Search
SVM	–	Support Vector Machine
tc	–	term count
a	–	step size
w	–	Pseudo-random integer
L	–	Lair
GA	–	Genetic Algorithm

CS	–	Cuckoo Search
RSS	–	Ringed Seal Search
RBF	–	Radial Basic Function

LIST OF PUBLICATIONS

1. Sharif, W., Samsudin, N. A., Deris, M. M., (2017). Improved Relative Discriminative Criterion Feature Ranking Technique for Text Classification. *International Journal of Artificial Intelligence*™, 15 (2), pp. 61-78. (ISI and Scopus Indexed)
2. Sharif, W., Samsudin, N. A., Deris, M. M., Khalid, S.K.A. (2018) Technical Study on Feature Ranking Techniques and Classification Algorithms, *International journal of engineering and applied sciences* Malaysia, 13(9), pp. 7074-7080. (Scopus Indexed)
3. Sharif, W., Samsudin, N. A., Deris, M. M. (2019). An optimized support vector machine with Ringed Seal Search algorithm for efficient text classification, *Journal of Engineering science and technology,* 14(3), pp. 1601-1613 (ISI and Scopus Indexed)

CHAPTER 1

INTRODUCTION

1.1. Research Background

A lot of digital information is associated with the World Wide Web and Internet, which is in both structured and unstructured form. In the early sixties of the 20th century, generation of excessive data was observed. Many organizations produced a huge amount of digital data on daily basis and additionally, people also post multiple documents online. This information belongs to digital libraries, blog repositories, online forums, news articles, and biological databases shown in Figure 1.1. Accurate classification of such data is not an easy task as it requires a lot of effort. However, machine learning helps to automatically analyses the data by identifying the patterns for making classification with minimal human intervention.

In machine learning, to extract useful information that is relevant to the interest from constantly increasing documents becomes a vital task. Documents can be in various formats such as word, phrase, term, pattern, concept, sentence, paragraph and text (Wang, Wang, & Wei, 2018; Wang & Zhang, 2013). This excessive information requires some proficient classification algorithms which can be used to assign documents into one or more classes (labels) (Elarnaoty & Farghaly, 2018; Sokolova, 2018). The classification algorithms are applied on different text applications such as sentiment analysis (Sharif *et al.,* 2016; Zheng, Wang, & Gao, 2018), text clustering (Jiang *et al.,* 2012; Sharif *et al.,* 2017; Su *et al.,* 2018; Uddin *et al.,* 2016), spam filtering (Midigo, Mwangi, & Okeyo, 2017), website classification (Tyagi & Gupta, 2018; Devi, 2008), disease report finding (Conway *et al.,* 2009), document summarization and text classification (Sharif *et al.,* 2018b; Su *et*

al., 2018).

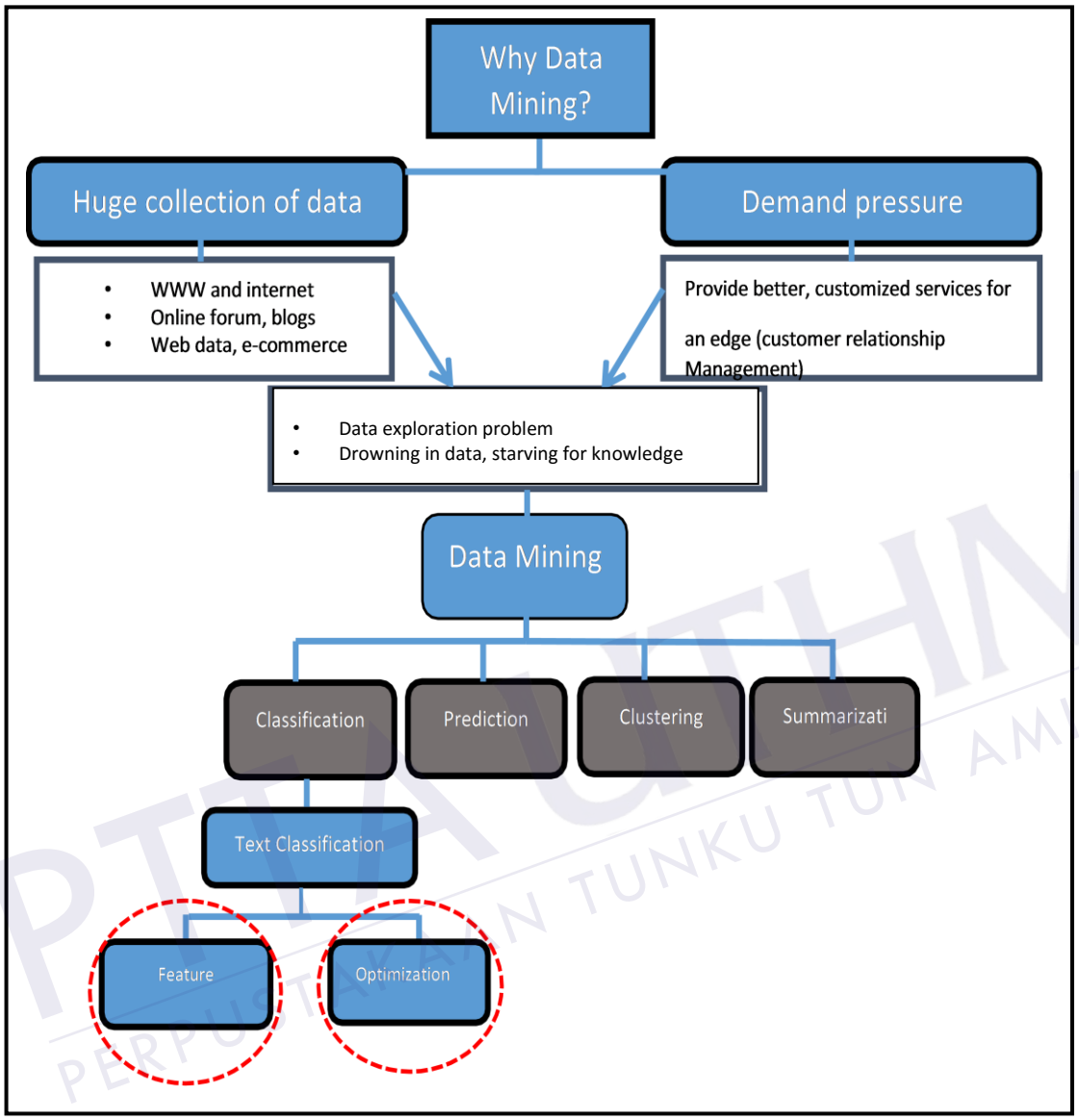

Figure1.1: Data mining for feature selection and classification

Text classification has become an active research area over the last decade. Past studies (Adeleke *et al*., 2018; Silva, Almeida, & Yamakami, 2017; Vijayan, Bindu, & Parameswaran, 2017) indicated that information retrieval plays an important role to improve accuracy in text classification. Textual data are highly dimensional that it must be pre-processed before applying classification algorithms (Fragoudis, Meretakis, & Likothanassis, 2005). Therefore, it takes much time to discover the knowledge of interest from textual data (Herawan, Yanto, & Deris, 2009; Yang *et al.*, 2012). The advent of high dimensional data has carried unprecedented challenges to machine learning researchers, making the learning task

more complex and computationally demanding. The term high dimensionality is applied to a database containing a huge number of features. On text dataset, there are two main tasks which affect the accuracy: feature selection and classification (Karaca & Bayir, 2017). Feature selection algorithm means how to select informative features from dataset. Feature selection algorithms select relevant features which produce significant result (Chen *et al.,* 2009; Venkataraman & Selvaraj, 2018). In classification process, feature selection is a task that minimizes the classification errors and improve the results (García *et al.,* 2016). However, feature selection is critical to solve the problem of handling high dimensional data (Parlak & Uysal, 2018; Pinheiro *et al.,* 2012).

Most text classification algorithms use vector space model and bag-of-words representation to model textual documents. In the vector space model, every word or group of words (depending on whether working with a single word or a phrase) is called a term, which represents one dimension of the feature space. A positive number, reflecting the relevancy and significance, is assigned to each term. This number becomes the frequency of the term in the document (Makrehchi, 2007). If words are selected as terms, dimensionality consists of the number of distinct words in vocabulary (Liu & Fu, 2012). In moderate text dataset, the number of words (terms) can be simply grown in tens of thousands which has irrelevant terms. Irrelevant and unwanted terms in text dataset produce noise that diminishes the performance of classification algorithms. To avoid these irrelevant and unwanted data, feature selection algorithms are used. The feature selection algorithms get useful terms that can improve the accuracy of the text classification algorithms (Karaca & Bayir, 2017; Li *et al.,* 2017; Yun *et al.,* 2017). Feature selection algorithms have the capability to keep discriminative features and reduced dimensionality (Zheng *et al.,* 2017; Liu *et al.,* 2018; Ludwig *et al.,* 2017). There are many types of feature selection algorithms as shown in Figure 1.2.

Feature selection methods are split up into feature subset selection and feature ranking. This division is based on how the features are combined for classification evaluation (Gnana, Appavu, & Leavline, 2016; Venkataraman & Selvaraj, 2018). In filter based methods, such as Information Gain (Lee & Lee, 2006), Chi-square (Manning, Raghavan, & Schütze, 2008), and Distinguishing Feature Selector (DFS) (Uysal & Gunal, 2012), individual features are used which produce a good result. These filter techniques reduce dimensionality which provides

less computation, reduces over fitting and increases generalization (Meenakshi, 2013).

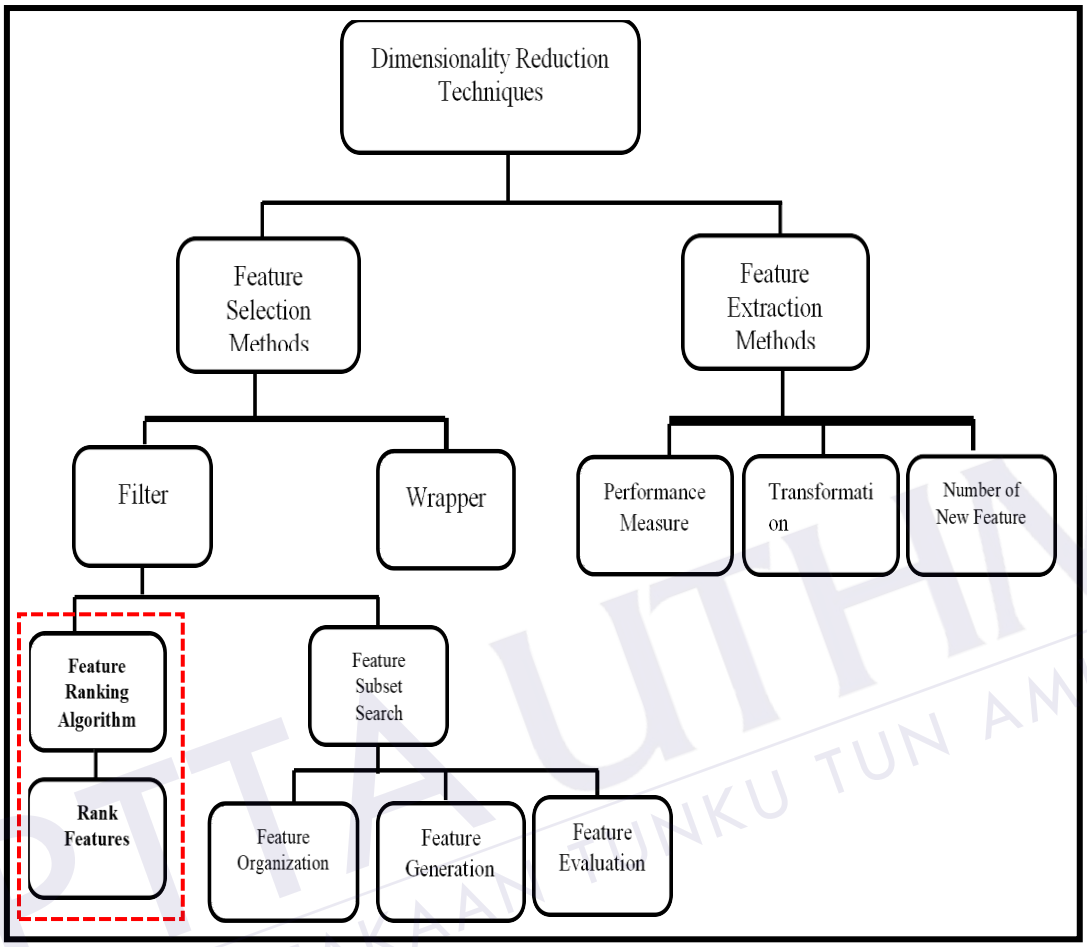

Figure1.2: Feature Selection (Khalid, Khalil, & Nasreen, 2014)

A few past studies (e.g., Rehman *et al.*, 2015; Wang *et al.*, 2015) have been carried to overcome the deficiency of the prior feature selection techniques (i.e., Distinguishing Feature Selector (DFS). DFS use the terms which occur frequently in one class. Baccianella *et al.* (2013) described that existing feature selection did not consider term frequencies (term counts) to rank a term. For considering the term frequencies, documents were logically break down into several micro-documents where each micro-document consists of only one word. They found that using micro- document improves classification accuracy of the textual data. Rehman *et al.* (2015) proposed Relative Discriminative Criterion (RDC) that splits the document frequency of a term into document frequencies for each term count and rank the final term. RDC considers the difference between document frequencies for respective

term counts of a term in the positive and negative classes. Subsequently, Wang *et al.* (2015) modified the RDC by normalizing the financial dataset because some financial documents are very long and some are very small. The long documents caused biased results in classification.

Feature selection techniques play an important role in enhancing the accuracy and reliability of the text classification algorithms (Rehman *et al.,* 2015). In text classification, Li and Chen (2012) illustrated that the higher dimensionality of feature space impose weighty overhead to build document classifier. There are features that can be unwanted or irrelevant which misguide the classification result and even make some classification algorithms such as Support Vector Machine (SVM) inapplicable (Chen *et al.,* 2009; Tang *et al.,* 2016). In terms of accuracy, computational time, and stability of parameter setting, SVM is much better (Kim *et al.,* 2002) as compared to other non-parametric classification techniques such as neural network, k-nearest neighbor (k-NN). However, carefully selected relevant features may lead to enhance the performance of the classification methods.

There are two major types of classification methods (Herawan *et al.,* 2009): supervised classification and unsupervised classification. In unsupervised classification, the system has no pre-defined classes and no external mechanism used while in supervised learning, class label and external mechanism (human feedback), are used for classification. In supervised classification, there are many classifiers such as naive bayes, decision tree, KNN, Neural Network and SVM that can be applied to text data (Biran & Cotton, 2017). Comparatively, SVM has greater implementation requirements to fulfill the theory, but it has been used to produce better results in many fields (Li & Kong, 2014). SVM is a large margin classifier that found a decision boundary between classes. In other words, it is a discriminative classifier defined by a separating hyperplane (Khan *et al.,* 2010).

A major challenge to SVM adoption for studying real-world problems is the selection of parameter. There are many parameters of SVM that can be tuned to improve the performance of classification (Song *et al.,* 2011). The generalization ability of SVM is dependent on selecting and optimizing of parameters which are challenging tasks to improve the classification accuracy. A different number of parameters are required to be optimized contingent on the usage of linear or non-linear SVM. In the case of linear SVM, only one parameter needs to be optimized which is the penalty parameter. Whereas using Radial Basic Function (RBF) kernel

in case of non-linear SVM, both γ and C parameter need to be optimized for better accuracy achievement (Sun *et al.,* 2018). These parameters are tuned with metaheuristic search technique.

Metaheuristic techniques perform a significant role in various optimization problems. Metaheuristic techniques copy natural phenomena that evolved over millions of years (Kuo *et al.,* 2018). The major advantage of metaheuristic technique is that it maintains good performance with dynamic changes. The metaheuristic techniques have been used in many fields to solve global optimization problems (Peng *et al.,* 2014). There are many metaheuristic techniques of optimization such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO) algorithm, Cuckoo Search (CS) and Ringed Seal Search Algorithm (RSS). They are global optimization algorithms used to find the best hyper-plane to classify documents accurately. GA algorithm uses operators inspired by genetic variation and natural selection (İlhan & Tezel, 2013). Whereas PSO was inspired by the fish and bird swarm intelligence (Aghdam & Heidari, 2015). It is subsidized with a flexible and well-proportioned mechanism to improve global exploration capability (Abido, 2002). On the other hand, CS was inspired by the brood intelligent behavior of some cuckoo species and based on levy flight. The CS strategy consists of laying eggs in other cuckoos' nests (Kanagarajan & Arumugam, 2018) while RSS is inspired by the natural behavior of the seal pup for choosing the best lair to escape predators. In contrary to GA, PSO, and CS, RSS is faster in finding the global optima over its homologs and in keeping balance between exploitation and exploration (Saadi *et al.,* 2016). These optimization algorithms have the capability to optimize the parameter of SVM (Phan *et al.,* 2017; Sayed *et al.,* 2018), resulting in increased classification accuracy compare to conventional SVM (Manurung *et al.,* 2017).

Based on the above discussion, it is cleared that the feature selection technique and optimizing the parameter of SVM play critical role in accurate text classification. Multiple SVMs with different parameters have to be computed in order to find SVM that produces better classification performance in text classification. Text classification is an important field for researchers to optimize the parameters of SVM with search algorithms for improving classification accuracy (Chen *et al.,* 2013; İlhan & Tezel, 2013). However, this study focuses on improving the existing feature ranking algorithm (i.e. RDC) and optimizing the parameters of SVM by integrating the optimization algorithm (i.e. RSS).

1.2. Research Motivation

A great amount of textual data in digital form is generated every day due to the development of computers/Internet and WWW. Textual data is highly dimensional data, it has irrelevant and unwanted features which are difficult to manage and maintain. There is a need for feature selection techniques to remove these irrelevant features. Feature ranking is important in the feature selection phase because it can increase the accuracy of the classifier. Most existing techniques focus on frequently occurring features in feature ranking techniques but rare and informative features are ignored. Due to feature ranking, machine learning algorithms become more scalable, reliable and accurate. On the other hand, in text classification, the parameter setting is also important for SVM classifier to improve accuracy. Metaheuristic techniques are used to optimize the SVM parameters. These optimization algorithms provide faster and cost-effective results to classify the data accurately (Gaspar, Carbonell, & Oliveira, 2012).

1.3. Research Questions

The research questions of this study are as follows:

i. How Relative Discriminative Criterion feature ranking technique can be improved for feature selection purpose?
ii. How the parameter of SVM can be tuned for text classification problems?
iii. How to evaluate the performance of the proposed IRDC and RSS-SVM in text classification problems?

1.4. Research Problem

Existing literature explored two key issues that affect the overall efficacy of classification; first is the lack of consideration of rare and informative terms in feature selection whereas the second issue is the effective searching and selection in parameter setting. It is observed that existing systems did not consider term frequencies (term counts) to rank a term except in the case of Relative

REFFERENCE

Abido, M. (2002). Optimal design of power-system stabilizers using particle swarm optimization. *IEEE transactions on energy conversion*, 17(3), 406-413.

Aida Husaini, N., Ghazali, R., & Tri Riyadi Yanto, I. (2018). Enhancing Modified Cuckoo Search Algorithm by Using MCMC Random Walk. *International Conference on Science in Information Technology,* (pp. 1-6).

Albishre, K., Albathan, M., & Li, Y. (2015). Effective 20 newsgroups dataset cleaning. *Paper presented at the 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT).*

Adachi, Y., Onimura, N., Yamashita, T., & Hirokawa, S. (2017). Classification of Imbalanced Documents by Feature Selection. *Paper presented at the Proceedings of the International Conference on Compute and Data Analysis.* (pp. 228-232).

Adeleke, A. O., Samsudin, N. A., Mustapha, A., & Nawi, N. M. (2018). A Group-Based Feature Selection Approach to Improve Classification of Holy Quran Verses. *Paper presented at the International Conference on Soft Computing and Data Mining*. 2, (pp. 282-297).

Ageev, M. S., & Dobrov, B. V. (2003). Support Vector Machine Parameter Optimization for Text Categorization Problems. *Paper presented at the ISTA.* (pp. 165-176).

Aghdam, M. H., & Heidari, S. (2015). Feature selection using particle swarm optimization in text categorization. *Journal of Artificial Intelligence and Soft Computing Research*, 5(4), 231-238.

Agnihotri, D., Verma, K., & Tripathi, P. (2016). Computing symmetrical strength of N-grams: a two pass filtering approach in automatic classification of text documents. *SpringerPlus*, 5(1), 942.

Agnihotri, D., Verma, K., & Tripathi, P. (2017). Variable Global Feature Selection Scheme for automatic classification of text documents. *Expert Systems with Applications*, 81, 268-281.

Al-Anzi, F. S., & AbuZeina, D. (2017). Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. *Journal of King Saud University-Computer and Information Sciences*, 29(2), 189-195.

Asir, D., Appavu, S., & Jebamalar, E. (2016). Literature Review on Feature Selection Methods for High-Dimensional Data. *International Journal of Computer Applications*, 136(1), 9-17.

Aurangabadkar, S., & Potey, M. A. (2014, February). Support Vector Machine based classification system for classification of sport articles. *International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*(pp. 146-150). IEEE.

Azam, N., & Yao, J. (2012). Comparison of term frequency and document frequency based feature selection metrics in text categorization. *Expert Systems with Applications*, 39(5), 4760-4768.

Baccianella, S., Esuli, A., & Sebastiani, F. (2013). Using micro-documents for feature selection: The case of ordinal text classification. *Expert Systems with Applications*, 40(11), 4687-4696.

Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1), 4-20.

Basu, T., & Murthy, C. A. (2012, December). Effective text classification by a supervised feature selection approach. *In IEEE 12th International Conference on Data Mining* (pp. 918-925).

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., & Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS Comput Biol*, 4(10), e1000173.

Bhan, I., Mosesso, K., Goyal, L., Philipp, J., Kalinich, M., Franses, J. W., Maheswaran, S. (2018). Detection and Analysis of Circulating Epithelial Cells in Liquid Biopsies From Patients With Liver Disease. *Gastroenterology*, 155(6), 2016-2018. e2011.

Bhatia, K., Jain, H., Kar, P., Varma, M., & Jain, P. (2015). Sparse local embeddings for extreme multi-label classification. *Paper presented at the Advances in*

Neural Information Processing Systems. (pp.730-738).

Bhushan, S. B., & Danti, A. (2017). Classification of text documents based on score level fusion approach. *Pattern Recognition Letters*, 94, 118-126.

Bijalwan, V., Kumar, V., Kumari, P., & Pascual, J. (2014). KNN based machine learning approach for text and document mining. *International Journal of Database Theory and Application*, 7(1), 61-70.

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. *Paper presented at the IJCAI-17 Workshop on Explainable AI (XAI).* Melbourne, Australia. 20 (8), (pp.8-13).

Bolón Canedo, V. (2014). *Novel feature selection methods for high dimensional data*. FACULTY OF INFORMATICS Department of Computer Science: Ph.D Thesis.

Bhushan, S. B., & Danti, A. (2017). Classification of text documents based on score level fusion approach. *Pattern Recognition Letters*, *94*, (p.118-126).

Bzdok, D., Krzywinski, M., & Altman, N. (2018). Machine learning: Supervised methods, SVM and kNN. *Nature Methods*, *Nature Publishing Group*, 2018, (pp.1-6).

Che, J., Yang, Y., Li, L., Bai, X., Zhang, S., & Deng, C. (2017). Maximum relevance minimum common redundancy feature selection for nonlinear data. *Information Sciences*, 409, 68-86.

Chen, H., Jiang, W., Li, C., & Li, R. (2013). A heuristic feature selection approach for text categorization by using chaos optimization and genetic algorithm. *Mathematical problems in Engineering*, 2013.

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, 36(3), 5432-5435.

Cherkassky, V., & Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural networks*, 17(1), 113-126.

Cristianini, N., Campbell, C., & Shawe-Taylor, J. (1999). Dynamically adapting kernels in support vector machines. *In Advances in neural information processing systems* (pp. 204-210).

Conway, M., (2009). "Classifying disease outbreak reports using n-grams and semantic features." *International journal of medical informatics* 78(12): e47-e58.

Civicioglu, P., & Besdok, E. (2013). A conceptual comparison of the Cuckoo-search,

particle swarm optimization, differential evolution and artificial bee colony algorithms. *Artificial intelligence review,* 39(4), 315-346.

Conneau, A., Schwenk, H., Barrault, L., & Lecun, Y. (2017). Very deep convolutional networks for text classification. *Paper presented at the Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics:* Volume 1, Long Papers.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

de Almeida, B. J., Neves, R. F., & Horta, N. (2018). Combining Support Vector Machine with Genetic Algorithms to optimize investments in Forex markets with high leverage. *Applied Soft Computing,* 64, 596-613.

Di Martino, S., Ferrucci, F., Gravino, C., & Sarro, F. (2011). A genetic algorithm to configure support vector machines for predicting fault-prone components. Paper presented at the *International Conference on Product Focused Software Process Improvement.* Springer, Berlin, Heidelberg. (pp. 247-261).

Dsouza, K. J., & Ansari, Z. A. (2015). A novel data mining approach for multi variant text classification. *Paper presented at the Cloud Computing in Emerging Markets (CCEM), IEEE International Conference on.* (pp. 68-73).

Ding, S., & Chen, L. (2010). Intelligent optimization methods for high-dimensional data classification for support vector machines. *Intelligent Information Management,* 2(06), 354.

Elarnaoty, M., & Farghaly, A. (2018). Machine Learning Implementations in Arabic Text Classification Intelligent Natural Language Processing: *Trends and Applications* (pp. 295-324).

Elhadad, M. K., Badran, K. M., & Salama, G. I. (2018). A Novel Approach for Ontology-Based Feature Vector Generation for Web Text Document Classification. *International Journal of Software Innovation (IJSI),* 6(1), pp.1- 10.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research,* 3, 1289-1305.

Fragoudis, D., Meretakis, D., & Likothanassis, S. (2005). Best terms: an efficient feature-selection algorithm for text categorization. *Knowledge and Information Systems,* 8(1), 16-33.

Fanjin, M., Ling, H., Jing, T., & Xinzheng, W. (2017). The Research of Semantic

Kernel in SVM for Chinese Text Classification. *Paper presented at the Proceedings of the 2nd International Conference on Intelligent Information Processing.*

Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review,* 47(1), 1-66.

García-Torres, M., Gómez-Vela, F., Melián-Batista, B., & Moreno-Vega, J. M. (2016). High-dimensional feature selection via feature grouping: A Variable Neighborhood Search approach. *Information Sciences,* 326, 102-118.

Gasanova, T. (2015). *Novel methods for text preprocessing and classification.* Ulm, Universität Ulm, Diss., 2015.

Gaspar, P., Carbonell, J., & Oliveira, J. L. (2012). On the parameter optimization of Support Vector Machines for binary classification. *Journal of Integrative Bioinformatics (JIB),* 9(3), 33-43.

Gnana, D. A. A., Appavu, S., & Leavline, E. J. (2016). Literature Review on Feature Selection Methods for High-Dimensional Data. *methods,* 136(1).

Garšva, G., & Danenas, P. (2014). Particle swarm optimization for linear support vector machines based classifier selection. Nonlinear Analysis: *Modelling and Control*, 19(1), 26-42.

Goldberg, D. E., & Holland, J. H. (1988). Genetic algorithms and machine learning. *Machine learning,* 3(2), 95-99.

Hancer, E., Xue, B., & Zhang, M. (2018). Differential evolution for filter feature selection based on information theory and feature ranking. *Knowledge-Based Systems,* 140, 103-119.

Herawan, T., Yanto, I. T. R., & Deris, M. M. (2009). Rough set approach for categorical data *clustering Database Theory and Application* (pp. 179-186): Springer.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *The annals of statistics,* pp.1171-1220.

Huang, Y. (2003). Support vector machines for text categorization based on latent semantic indexing. Electrical and Computer Engineering Department, *The Johns Hopkins University, Tech. Rep*.

Huerta, E. B., Duval, B., & Hao, J.-K. (2006). A hybrid GA/SVM approach for gene selection and classification of microarray data. *Paper presented at the Workshops on Applications of Evolutionary Computation.*

Huang, H., Qian, L., & Wang, Y. (2012). A SVM-based technique to detect phishing URLs. *Information Technology Journal*, 11(7), 921-925.

Iiritano, S., & Ruffolo, M. (2001). Managing the knowledge contained in electronic documents: a clustering method for text mining. *Paper presented at the Database and Expert Systems Applications, Proceedings. 12th International Workshop on.* (pp. 454-458).

İlhan, İ., & Tezel, G. (2013). A genetic algorithm–support vector machine method with parameter optimization for selecting the tag SNPs. *Journal of biomedical informatics*, 46(2), 328-340.

Jiang, J.-Y., Liou, R.-J., & Lee, S.-J. (2011). A fuzzy self-constructing feature clustering algorithm for text classification. *IEEE transactions on knowledge and data engineering*, 23(3), 335-349.

Jiang, L., Li, C., Wang, S., & Zhang, L. (2016). Deep feature weighting for naive Bayes and its application to text classification. *Engineering Applications of Artificial Intelligence*, 52, 26-39.

Joachims, T. (1998, April). Text categorization with support vector machines: Learning with many relevant features. *In European conference on machine learning* (pp. 137-142). Springer, Berlin, Heidelberg.

Kanagarajan, K., & Arumugam, S. (2018). Intelligent sentence retrieval using semantic word based answer generation algorithm with cuckoo search optimization. *Cluster Computing*, pp.1-11.

Kumbhar, P., & Mali, M. (2016). A survey on feature selection techniques and classification algorithms for efficient text classification. *International Journal of Science and Research*, 5(5), 9.

Karaca, M. F., & Bayir, S. (2017). Examining the Impact of Feature Selection Methods on Text Classification. *International journal of advanced computer science and applications*, 8(12), 380-388.

Karegowda, A. G., Manjunath, A., & Jayaram, M. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2), 271-277.

Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. *Paper presented at the Science and Information Conference (SAI)* (pp. 372-378).

Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1), 4-20.

Kim, H., Howland, P., & Park, H. (2005). Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research*, 6(Jan), 37-53.

Kim, K. I., Jung, K., Park, S. H., & Kim, H. J. (2002). Support vector machines for texture classification. *IEEE Transactions on pattern analysis and machine intelligence*, 24(11), 1542-1550.

Ko, Y., Park, J., & Seo, J. (2002). Automatic text categorization using the importance of sentences. *Paper presented at the Proceedings of the 19th international conference on Computational linguistics*-1(8), (pp. 1-7)

Kulkarni, A. R., Tokekar, V., & Kulkarni, P. (2012). Identifying context of text documents using Naïve Bayes classification and Apriori association rule mining, *In Software Engineering (CONSEG), CSI Sixth International Conference on*, (pp. 1-4).

Kirchner, A., & Signorino, C. S. (2018). Using Support Vector Machines for Survey Research. *Survey Practice*, 11(1), 2715.

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, 114, 128-147.

Kuo, R., Huang, S. L., Zulvia, F., & Liao, T. (2018). Artificial bee colony-based support vector machines with feature selection and parameter optimization for rule extraction. *Knowledge and Information Systems*, 55(1), 253-274.

Koch, P. (2013). *Efficient tuning in supervised machine learning.* Leiden Institute of Advanced Computer Science (LIACS), Faculty of Science, Leiden University, Doctoral dissertation.

Labani, M., Moradi, P., Ahmadizar, F., & Jalili, M. (2018). A novel multivariate filter method for feature selection in text classification problems. *Engineering Applications of Artificial Intelligence*, 70, 25-37.

LaValle, S. M., Branicky, M. S., & Lindemann, S. R. (2004). On the relationship between classical grid search and probabilistic roadmaps. *The International Journal of Robotics Research*, 23(7-8), 673-692.

Law, M. (2011). *A simple introduction to support vector machines.* Lecture for CSE, 802. Retrieved from http://www.cise.ufl.edu/class/cis4930sp11dtm/n

otes/intro_svm_new.pdf

Lee, C., & Lee, G. G. (2006). Information gain and divergence-based feature selection for machine learning-based text categorization. *Information Processing & Management*, 42(1), 155-165.

Li, Q., Gu, Y., & Jia, J. (2017). Classification of multiple Chinese liquors by means of a QCM-based e-nose and MDS-SVM classifier. *Sensors*, 17(2), 272.

Li, B. (2016). Importance weighted feature selection strategy for text classification. *Paper presented at the Asian Language Processing (IALP), International Conference on*.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6), 94.

Li, X., & Kong, J. (2014). Application of GA–SVM method with parameter optimization for landslide development prediction. *Natural Hazards and Earth System Sciences*, 14(3), 525-533.

Li, Y., & Chen, C. (2012). Research on the feature selection techniques used in text classification. *Paper presented at the Fuzzy Systems and Knowledge Discovery (FSKD), 9th International Conference on.* (pp. 725-729).

Li, Y., Li, T., & Liu, H. (2017). Recent advances in feature selection and its applications. *Knowledge and Information Systems*, 53(3), 551-577.

Li, B. (2016). Importance weighted feature selection strategy for text classification. *Paper presented at the Asian Language Processing (IALP), International Conference on.* (pp. 344-347).

Liu, C., He, L., Li, Z., & Li, J. (2018). Feature-Driven Active Learning for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(1), 341-354.

Liu, T.-Y., Xiubo, G., Qin, T., & Li, H. (2010). *Feature selection for ranking:* Google Patents.

Liu, X., & Fu, H. (2012). *A hybrid algorithm for text classification problem.* Przeglad Elektrotechniczny, 88(1B), 8-11.

Ludwig, A. R., Piorek, H., Kelch, A. H., Rex, D., Koitka, S., & Friedrich, C. M. (2017). *Improving model performance for plant image classification with filtered noisy images.* Working Notes of CLEF, 2017.

Makrehchi, M. (2007). *Feature ranking for text classifiers.* Electrical and Computer

Engineering, University of Waterloo, Ph.D Thesis.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval (Vol. 1): *Cambridge university press Cambridge.*

Martín, R., Aler, R., & Galván, I. M. (2018). A filter attribute selection method based on local reliable information. *Applied Intelligence,* 48(1), 35-45.

Masood, M. K., Jiang, C., & Soh, Y. C. (2018). A novel feature selection framework with Hybrid Feature-Scaled Extreme Learning Machine (HFS-ELM) for indoor occupancy estimation. *Energy and Buildings,* 158, 1139-1151.

Manikandan, R., & Sivakumar, R. (2018). Machine learning algorithms for text-documents classification: *International Journal of Academic Research and Development, 3(2).*

Meenakshi.R, V. S. (2013). Structured data extraction from the deep web. *IOSR Journal of Computer Engineering* (IOSR-JCE), 8(6), 93-96.

Midigo, R. O., Mwangi, W., & Okeyo, G. O. (2017). Biterm for spam filtering in short message service text. *International Journal of Computer Science Issues (IJCSI),* 14(1), 79.

Mladenić, D., Brank, J., Grobelnik, M., & Milic-Frayling, N. (2004). Feature selection using linear classifier weights: interaction with classification models. *In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 234-241).

Mladenic, D., & Grobelnik, M. (1999). Feature selection for unbalanced class distribution and naive bayes. *Paper presented at the ICML*. Vol. 99(6), pp. (258-267).

Mocherla, S., Danehy, A., & Impey, C. (2018). Evaluation of Naive Bayes and Support Vector Machines for Wikipedia. *Applied Artificial Intelligence,* 1-12.

Morimoto, M., Shirai, Y., Takeda, H., & Hasuko, K. (2017). *Document classification system, document classification method, and document classification program:* Google Patents.

Moschitti, A., & Basili, R. (2004). Complex linguistic features for text classification: A comprehensive study Advances in Information Retrieval (pp. 181-196): *Springer.*

Monika, L., (2016). Variable Features Selection for Classification of Medical Data using SVM. *International Journal of Engineering Technology, Management and Applied Sciences*,4(5).

Manurung, J., Mawengkang, H., & Zamzami, E. (2017). Optimizing Support Vector Machine Parameters with Genetic Algorithm for Credit Risk Assessment. *Paper presented at the Journal of Physics: Conference Series.*

Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006). Spam filtering with naive bayes-which naive bayes? *Paper presented at the CEAS.*

Mohapatra, P., Chakravarty, S., & Dash, P. K. (2015). An improved cuckoo search based extreme learning machine for medical data classification. *Swarm and Evolutionary Computation,* 24, 25-49.

Min, S. H., & Han, I. (2005, July). Recommender systems using support vector machines. *In International Conference on Web Engineering*. Springer, Berlin, Heidelberg, (pp. 387-393).

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications,* 41(16), 7653-7670.

Novaković, J. (2016). Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research,* 21(1).

Nayak, J., Naik, B., & Behera, H. (2015). A comprehensive survey on support vector machine in data mining tasks: applications & challenges. *International Journal of Database Theory and Application,* 8(1), 169-186.

Ng, A. Y. (2004, July). Feature selection, L 1 vs. L 2 regularization, and rotational invariance. *In Proceedings of the twenty-first international conference on Machine learning, ACM* (p. 78).

Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications,* 57, pp. 232-247.

Özgür, L., & Güngör, T. (2010). Text classification with the support of pruned dependency patterns. *Pattern Recognition Letters,* 31(12), pp.1598-1607.

Özgür, A., & Güngör, T. (2006, June). Classification of skewed and homogenous document corpora with class-based and corpus-based keywords. *In Annual Conference on Artificial Intelligence* (pp. 91-101). Springer, Berlin, Heidelberg.

Parlak, B., & Uysal, A. K. (2016). The impact of feature selection on medical document classification. *Paper presented at the Information Systems and Technologies (CISTI), 11th Iberian Conference on*. (pp. 1-5).

Parlak, B., & Uysal, A. K. (2018). On Feature Weighting and Selection for Medical Document Classification. *Developments and Advances in Intelligent Systems and Applications* (pp. 269-282).

Paul, A. (2014). *Effect of imbalanced data on document classification algorithms.* Auckland University of Technology. Ph.D Thesis.

Patel, A. D., & Sharma, Y. K. (2019). Web Page Classification on News Feeds Using Hybrid Technique for Extraction. *Information and Communication Technology for Intelligent Systems* (pp. 399-405): Springer.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence,* 27(8), 1226-1238.

Peng, H. X., Du, M. X., Wang, B., & Wang, P. (2014). Acquisition of Swarm Intellegence (SI) Principle in Practice Teaching of Automation Disciplines. *Paper presented at the Advanced Materials Research*. Vol. 926, (pp. 4443-4446).

Phan, A. V., Le Nguyen, M., & Bui, L. T. (2017). Feature weighting and SVM parameters optimization based on genetic algorithms for classification problems. *Applied Intelligence,* 46(2), 455-469.

Pinheiro, R. H., Cavalcanti, G. D., Correa, R. F., & Ren, T. I. (2012). A global-ranking local feature selection method for text categorization. *Expert Systems with Applications*, 39(17), 12851-12857.

Precup, R.-E., Preitl, S., & Korondi, P. (2007). Fuzzy controllers with maximum sensitivity for servosystems. *IEEE Transactions on Industrial Electronics,* 54(3), 1298-1310.

Pillai, N., & Matuszek, C. (2018). Unsupervised selection of negative examples for grounded language learning. *Paper presented at the Thirty-Second AAAI Conference on Artificial Intelligence.*

Raju, M. K., Subrahmanian, S. T., & Sivakumar, T. (2017). A Comparative Survey on Different Text Categorization Techniques. *International Journal of Computer Science and Engineering*, 5(3), 1612-1618.

Rehman, A., Javed, K., Babri, H. A., & Saeed, M. (2015). Relative discrimination criterion–a novel feature ranking method for text data. *Expert systems with Applications*, 42(7), 3670-3681.

Ruchika Singh Rajput, Agrawal. j, Sanjeev, (2017) Cuckoo Search based Hybrid Classification Techniques – *international journal of computer science and technology, ijcst,* 8(1)

Rodriguez-Galiano, V., Luque-Espinar, J., Chica-Olmo, M., & Mendes, M. (2018). Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. *Science of The Total Environment,* 624, 661-672.

Ren, Y., & Bai, G. (2010). Determination of optimal SVM parameters by using GA/PSO. *JCP*, 5(8), 1160-1168.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19), 2507-2517.

Salton, G., & McGill, M. (1983). Introduction to modern information Philadelphia, PA. American Association for Artificial Intelligence retrieval: *McGraw-Hil*l.[a] RA.

Saqib, S. M., Kundi, F. M., & Ahmad, S. (2018). Unsupervised Learning Method for Sorting Positive and Negative Reviews Using LSI (Latent Semantic Indexing) with Automatic Generated Queries. *IJCSNS*, 18(1), 56.

Sayed, G. I., Soliman, M., & Hassanien, A. E. (2018). Modified Optimal Foraging Algorithm for Parameters Optimization of Support Vector Machine. *Paper presented at the International Conference on Advanced Machine Learning Technologies and Applications.* (pp. 23-32).

Sathiaseelan, J. (2015). A technical study on Information Retrieval using web mining techniques. *Paper presented at the Innovations in Information, Embedded and Communication Systems (ICIIECS), International Conference on.* (pp. 1-5).

Saxena, A. K., & Dubey, V. K. (2015). A survey on feature selection algorithms. *Int. J. Recent Innov. Trends Comput. Commun,* 3, 1895-1899.

Scott, S., & Matwin, S. (1999). Feature engineering for text classification. *Paper presented at the ICML.* (Vol. 99, pp. 379-388).

Sebastiani, F. (2005). *Text Categorization,* Department of the Mathematic Pura e Applicata University the Padova, Italy.

Senan, N., Ibrahim, R., Nawi, N. M., Yanto, I. T. R., & Herawan, T. (2011). Rough set approach for attributes selection of traditional Malay musical instruments sounds classification. *Paper presented at the International Conference on*

Ubiquitous Computing and Multimedia Applications.

Shang, J., Liu, J., Jiang, M., Ren, X., Voss, C. R., & Han, J. (2018). Automated phrase mining from massive text corpora. *IEEE transactions on knowledge and data engineering.* 30(10), pp.1825-1837.

Sharif, W., Samsudin, N. A., Deris, M. M., & Aamir, M. (2017). Improved Relative Discriminative Criterion Feature Ranking Technique for Text Classification. *International Journal of Artificial Intelligence™, 1*5(2), 61-78.

Sharif, W., Samsudin, N. A., Deris, M. M., & Naseem, R. (2016). Effect of negation in sentiment analysis. *Paper presented at the Innovative Computing Technology (INTECH), Sixth International Conference on*. (pp. 718-723).

Sharif, W., Samsudin, N. A., Deris, M. M., & SKA, Khalid (2018). A Technical Study on Feature Ranking Techniques and Classification Algorithms, *Journal of Engineering and Applied Sciences* 13(9), 7074-7080.

Sharif, W., Samsudin, N. A., Deris, M. M. (2019). An optimized support vector machine with Ringed Seal Search algorithm for efficient text classification, *Journal of Engineering science and technology*, 14(3), pp. 1601-1613.

Silva, R. M., Almeida, T. A., & Yamakami, A. (2017). MDLText: An efficient and lightweight text classifier. *Knowledge-Based Systems*, 118, pp. 152-164.

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. *Paper presented at the Computing for Sustainable Global Development (INDIACom), 3rd International Conference on*. (pp. 1310-1315).

Singh, J., & Gupta, V. (2017). A systematic review of text stemming techniques. *Artificial Intelligence Review*, 48(2), pp.157-217.

Siolas, G., & d'Alché-Buc, F. (2000). Support vector machines based on a semantic kernel for text categorization. *Paper presented at the Neural Networks, 2000. IJCNN, Proceedings of the IEEE-INNS-ENNS International Joint Conference on.* (Vol. 5, pp. 205-209).

Smith, A. M., Jacobson, J. R. R., Kobashikawa, B. T., & Thatcher, G. G. (2018). *Spam filtering and person profiles:* Google Patents.

Sokolova, M. (2018). Big Text advantages and challenges: classification perspective. *International Journal of Data Science and Analytics*, 5(1), 1-10.

Song, H., Xue, Y., & ZHANG, L.-j. (2011). Research on kernel function selection simulation based on SVM classification [J]. *Computer and Modernization*, 8,

133-136.

Su, Y., Huang, Y., & Kuo, C.-C. J. (2018). Efficient Text Classification Using Tree-structured Multi-linear Principle Component Analysis. *arXiv preprint arXiv*:1801.06607.

Sun, L., Bao, J., Chen, Y., & Yang, M. (2018). Research on parameter selection method for support vector machines. *Applied Intelligence*, 1-12.

Saadi, Y., et al. (2016). "Ringed Seal Search for Global Optimization via a Sensitive Search Model." *PloS one* 11(1): e0144371.

Syarif, I., Prugel-Bennett, A., & Wills, G. (2016). SVM parameter optimization using grid search and genetic algorithm to improve classification performance. *Telkomnika*, 14(4), 1502.

Tan, F. (2007). Improving feature selection techniques for machine learning.

Al-Tahrawi, M. M. (2013). The role of rare terms in enhancing the performance of polynomial networks based text categorization. *Journal of Intelligent Learning Systems and Applications*, 5(02), 84.

Al-Tahrawi, M. M. (2014). The significance of low frequent terms in text classification. *International Journal of Intelligent Systems*, 29(5), 389-406.

Tang, B., He, H., Baggenstoss, P. M., & Kay, S. (2016). A Bayesian classification approach using class-specific features for text categorization. *IEEE transactions on knowledge and data engineering*, 28(6), 1602-1606.

Tan, P., Steinbach, M., & Kumar, V. *Introduction to data mining* (2014). Edimburgh Gate: Pearson Education Limited.

Tellez, E. S., Moctezuma, D., Miranda-Jímenez, S., & Graff, M. (2017). An Automated Text Categorization Framework based on Hyperparameter Optimization. *arXiv preprint* arXiv:1704.01975.

Tuba, E., Mrkela, L., & Tuba, M. (2016). Support vector machine parameter tuning using firefly algorithm. *Paper presented at the Radioelektronika (RADIOELEKTRONIKA), 26th International Conference.* (pp. 413-418).

Tyagi, N., & Gupta, S. K. (2018). Web Structure Mining Algorithms: A Survey Big Data Analytics (pp. 305-317): *Springer.*

Uddin, J., Ghazali, R., Deris, M. M., Naseem, R., & Shah, H. (2016). A survey on bug prioritization. *Artificial Intelligence Review*, 1-36.

Uysal, A. K. (2016). An improved global feature selection scheme for text classification. *Expert systems with Applications*, 43, 82-92.

Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems*, 36, 226-235.

Venkataraman, S., & Selvaraj, R. (2018). Optimal and Novel Hybrid Feature Selection Framework for Effective Data Classification Advances in Systems, Control and Automation (pp. 499-514): *Springer*.

Vijayan, V. K., Bindu, K., & Parameswaran, L. (2017). A comprehensive study of text classification algorithms. *Advances in Computing, Communications and Informatics, International Conference on Chicago*, (pp. 1109-1113).

Vergara, J. R., & Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural computing and applications*, 24(1), 175-186.

Wah, Y. B., Ibrahim, N., Hamid, H. A., Abdul-Rahman, S., & Fong, S. (2018). Feature Selection Methods: Case of Filter and Wrapper Approaches for Maximising Classification Accuracy. *Pertanika Journal of Science and Technology*, 26(1), 329-340.

Wajeed, M. A., & Adilakshmi, T. (2016). Adopting ant colony optimization for supervised text classification. *Paper presented at the Advances in Computing,Communications and Informatics (ICACCI), International Conference on*. (pp. 2562-2566).

Wang, B., Wang, L., & Wei, Q. (2018). TextZoo, a New Benchmark for Reconsidering Text Classification. *arXiv preprint arXiv*:1802.03656.

Wang, D., & Zhang, H. (2013). Inverse-category-frequency based supervised term weighting schemes for text categorization. *Journal of Information Science and Engineering*, 29(2), 209-225.

Wang, D., Zhang, H., Liu, R., Lv, W., & Wang, D. (2014). T-test feature selection approach based on term frequency for text categorization. *Pattern Recognition Letters*, 45, 1-10.

Wang, F., Zhang, Y., Xiao, H., Kuang, L., & Lai, Y. (2015). Enhancing Stock Price Prediction with a Hybrid Approach Based Extreme Learning Machine. *Paper presented at the Data Mining Workshop (ICDMW), IEEE International Conference on.* (pp. 1568-1575).

Wang, H., & Hong, M. (2017). Probability and Variance Score: an Efficient Supervised Feature Selection Method for Text Classification. *Journal of Residuals Science & Technology*, 14(3).

Wang, H., & Liu, S. (2016). An Effective Feature Selection Approach Using the

Hybrid Filter Wrapper. *International Journal of Hybrid Information Technology*, 9(1), 119-128.

Wang, X., Wang, J., Yang, Y., & Duan, J. (2017). Labeled LDA-Kernel SVM: A Short Chinese Text Supervised Classification Based on Sina Weibo. *Paper presented at the Information Science and Control Engineering (ICISCE), 4th International Conference on.* (pp. 428-432).

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., & Vapnik, V. (2001). Feature selection for SVMs. *Paper presented at the Advances in neural information processing systems*. (pp. 668-674).

Weina Niu ; Xiaosong Zhang ; Guowu Yang ; Zhiyuan Ma ; Zhongliu Zhuo (2018), Phishing Emails Detection Using CS-SVM, *International Conference on Ubiquitous Computing and Communications (ISPA/IUCC),* Guangzhou, China.

Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2016). A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation*, 20(4), 606-626.

Yang, H., Cui, H., & Tang, H. (2017). A text classification algorithm based on feature weighting. *Paper presented at the AIP Conference Proceedings*. Vol. 1864(8).

Yang, J., Liu, Y., Zhu, X., Liu, Z., & Zhang, X. (2012). A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. *Information Processing & Management,* 48(4), pp. 741-754.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *Paper presented at the ICML*. (Vol. 97(7), pp. 412-420).

Zhang, W., Tang, X., & Yoshida, T. (2015). TESC: An approach to TExt classification using Semi-supervised Clustering. *Knowledge-Based Systems,* 75, 152-160.

Zhefua, Y., Huibiaoa, L., & Chuanyingb, J. (2011). A fast optimization method for hybrid kernel parameters [J]. *Journal of Dalian Maritime University,* 3, 025.

Zheng, L., Wang, H., & Gao, S. (2018). Sentimental feature selection for sentiment analysis of Chinese online reviews. *International journal of machine learning and cybernetics,* 9(1), 75-84.

Zheng, W., Tang, D., Zhang, H., & Tang, H. (2017). Feature Selection with Structural Sparse Mode for Text Categorization. *Paper presented at the Intelligent*

Human-Machine Systems and Cybernetics (IHMSC), 9th International Conference on. (Vol. 1(8), pp. 359-362).